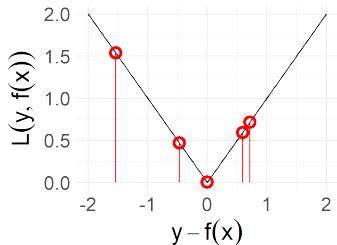


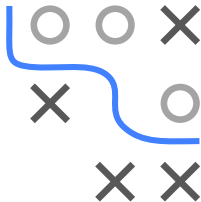
Introduction to Machine Learning

L1 Risk Minimizer (Deep-Dive)



Learning goals

- Derive the risk minimizer of the L1-loss
- Derive the optimal constant model for the L1-loss



L1-LOSS: RISK MINIMIZER

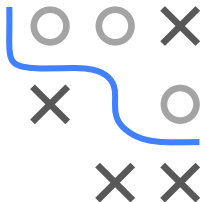
Proof: Let $p(y)$ be the density function of y . Then:

$$\begin{aligned}\arg \min_c \mathbb{E}[|y - c|] &= \arg \min_c \int_{-\infty}^{\infty} |y - c| p(y) dy \\ &= \arg \min_c \int_{-\infty}^c -(y - c) p(y) dy + \int_c^{\infty} (y - c) p(y) dy\end{aligned}$$

We now compute the derivative of the above term and set it to 0

$$\begin{aligned}0 &= \frac{\partial}{\partial c} \left(\int_{-\infty}^c -(y - c) p(y) dy + \int_c^{\infty} (y - c) p(y) dy \right) \\ &\stackrel{\text{*Leibniz}}{=} \int_{-\infty}^c p(y) dy - \int_c^{\infty} p(y) dy = \mathbb{P}_y(y \leq c) - (1 - \mathbb{P}_y(y \leq c)) \\ &= 2 \cdot \mathbb{P}_y(y \leq c) - 1 \\ \Leftrightarrow 0.5 &= \mathbb{P}_y(y \leq c),\end{aligned}$$

which yields $c = \text{med}_y(y)$.



L1-LOSS: RISK MINIMIZER

* **Note** that since we are computing the derivative w.r.t. the integration boundaries, we need to use Leibniz integration rule

$$\begin{aligned}\frac{\partial}{\partial c} \left(\int_a^c g(c, y) dy \right) &= g(c, c) + \int_a^c \frac{\partial}{\partial c} g(c, y) dy \\ \frac{\partial}{\partial c} \left(\int_c^a g(c, y) dy \right) &= -g(c, c) + \int_c^a \frac{\partial}{\partial c} g(c, y) dy\end{aligned}$$

We get

$$\begin{aligned}& \frac{\partial}{\partial c} \left(\int_{-\infty}^c -(y - c) p(y) dy + \int_c^{\infty} (y - c) p(y) dy \right) \\&= \frac{\partial}{\partial c} \left(\int_{-\infty}^c \underbrace{-(y - c) p(y)}_{g_1(c, y)} dy \right) + \frac{\partial}{\partial c} \left(\int_c^{\infty} \underbrace{(y - c) p(y)}_{g_2(c, y)} dy \right) \\&= \underbrace{g_1(c, c)}_{=0} + \int_{-\infty}^c \frac{\partial}{\partial c} (-(y - c)) p(y) dy - \underbrace{g_2(c, c)}_{=0} + \int_c^{\infty} \frac{\partial}{\partial c} (y - c) p(y) dy \\&= \int_{-\infty}^c p(y) dy + \int_c^{\infty} -p(y) dy.\end{aligned}$$



L1-LOSS: OPTIMAL CONSTANT MODEL

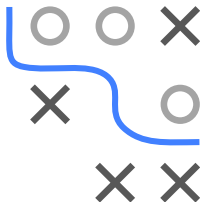
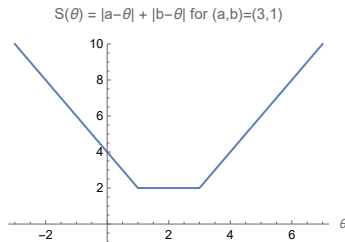
Proof:

- Firstly note that for $n = 1$ the median $\hat{\theta} = \text{med}(y^{(i)}) = y^{(1)}$ obviously minimizes the emp. risk \mathcal{R}_{emp} using the $L1$ loss.
- Hence let $n > 1$ in the following For $a, b \in \mathbb{R}$, define

$$S_{a,b} : \mathbb{R} \rightarrow \mathbb{R}_0^+, \theta \mapsto |a - \theta| + |b - \theta|$$

Any $\hat{\theta} \in [a, b]$ minimizes $S_{a,b}(\theta)$, because it holds that

$$S_{a,b}(\theta) = \begin{cases} |a - b|, & \text{for } \theta \in [a, b] \\ |a - b| + 2 \cdot \min\{|a - \theta|, |b - \theta|\}, & \text{otherwise.} \end{cases}$$



L1-LOSS: OPTIMAL CONSTANT MODEL

W.l.o.g. assume now that all $y^{(i)}$ are sorted in increasing order.

Let us define $i_{\max} = n/2$ for n even and $i_{\max} = (n-1)/2$ for n odd and consider the intervals

$$\mathcal{I}_i := [y^{(i)}, y^{(n+1-i)}], i \in \{1, \dots, i_{\max}\}.$$

By construction $\mathcal{I}_{j+1} \subseteq \mathcal{I}_j$ for $j \in \{1, \dots, i_{\max} - 1\}$ and $\mathcal{I}_{i_{\max}} \subseteq \mathcal{I}_i$.
With this, \mathcal{R}_{emp} can be expressed as

$$\begin{aligned}\mathcal{R}_{\text{emp}}(\theta) &= \sum_{i=1}^n L(y^{(i)}, \theta) = \sum_{i=1}^n |y^{(i)} - \theta| \\&= \underbrace{|y^{(1)} - \theta| + |y^{(n)} - \theta|}_{=S_{y^{(1)}, y^{(n)}}(\theta)} + \underbrace{|y^{(2)} - \theta| + |y^{(n-1)} - \theta|}_{=S_{y^{(2)}, y^{(n-1)}}(\theta)} + \dots \\&= \begin{cases} \sum_{i=1}^{i_{\max}} S_{y^{(i)}, y^{(n+1-i)}}(\theta) & \text{for } n \text{ is even} \\ \sum_{i=1}^{i_{\max}} (S_{y^{(i)}, y^{(n+1-i)}}(\theta)) + |y^{((n+1)/2)} - \theta| & \text{for } n \text{ is odd.} \end{cases}\end{aligned}$$



L1-LOSS: OPTIMAL CONSTANT MODEL

From this follows that

- for “ n is even”: $\hat{\theta} \in \mathcal{I}_{i_{\max}} = [y^{(n/2)}, y^{(n/2+1)}]$ minimizes S_i for all $i \in \{1, \dots, i_{\max}\} \Rightarrow$ it minimizes \mathcal{R}_{emp} ,
- for “ n is odd”: $\hat{\theta} = y^{(n+1)/2} \in \mathcal{I}_{i_{\max}}$ minimizes S_i for all $i \in \{1, \dots, i_{\max}\}$ and it's minimal for $|y^{((n+1)/2)} - \theta| \Rightarrow$ it minimizes \mathcal{R}_{emp} .

Since the median fulfills these conditions, we can conclude that it minimizes the $L1$ loss.

